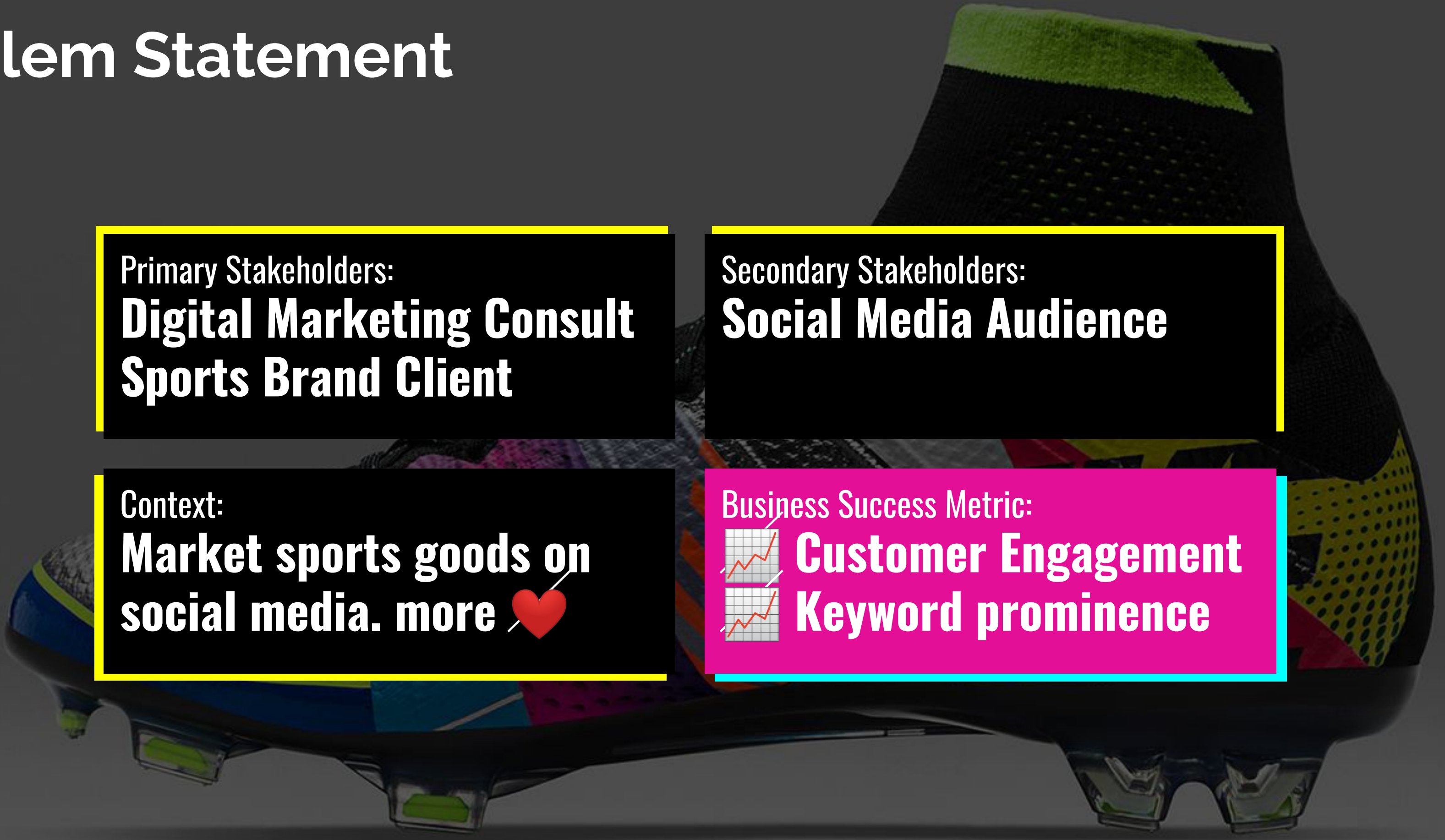
# Project 3:

Investment Opportunities in Distinguishing Billion Dollar Sports League Sentiments

By: Ivan, Yuan, John, Yun Jie

# Problem Statement



Primary Stakeholders:

**Digital Marketing Consult Sports Brand Client**

Secondary Stakeholders:

**Social Media Audience**

Context:

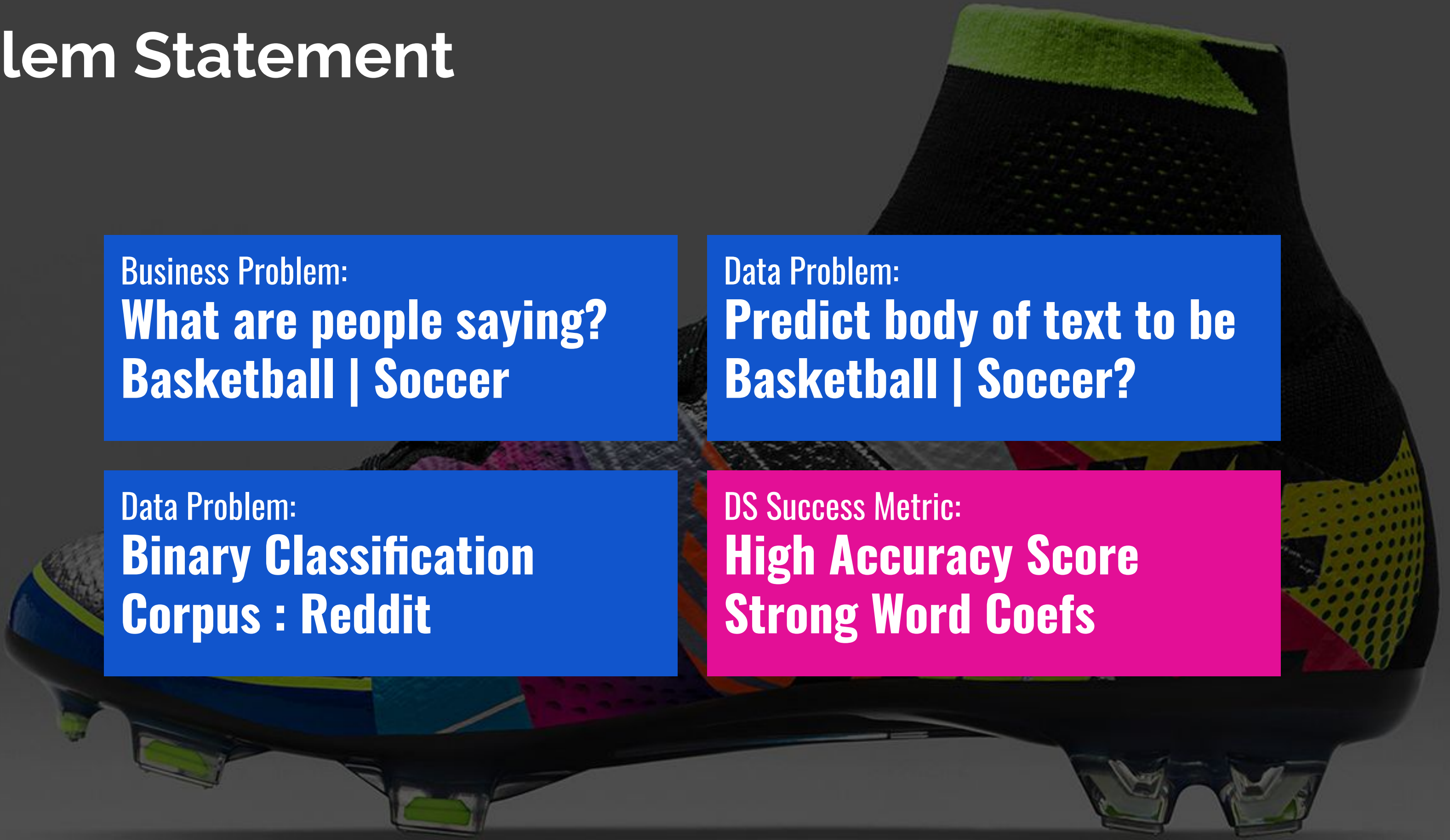
**Market sports goods on social media. more ❤️**

Business Success Metric:

**📈 Customer Engagement**

**📈 Keyword prominence**

# Problem Statement



Business Problem:

**What are people saying?**
**Basketball | Soccer**

Data Problem:

**Predict body of text to be Basketball | Soccer?**

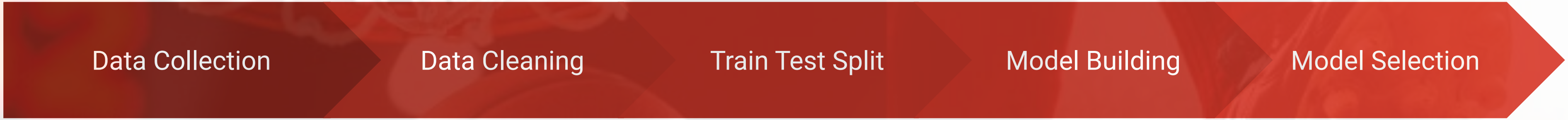
Data Problem:

**Binary Classification**
**Corpus : Reddit**

DS Success Metric:

**High Accuracy Score**
**Strong Word Coefs**

# Methodology



## Data Collection

Data was collected from:

- r/Basketball
- r/soccer
- Combined dataframe

## Data Cleaning

- Outliers removed
- Duplicated post removed
- Text cleaning
- Lemmatized

## Train Test Split

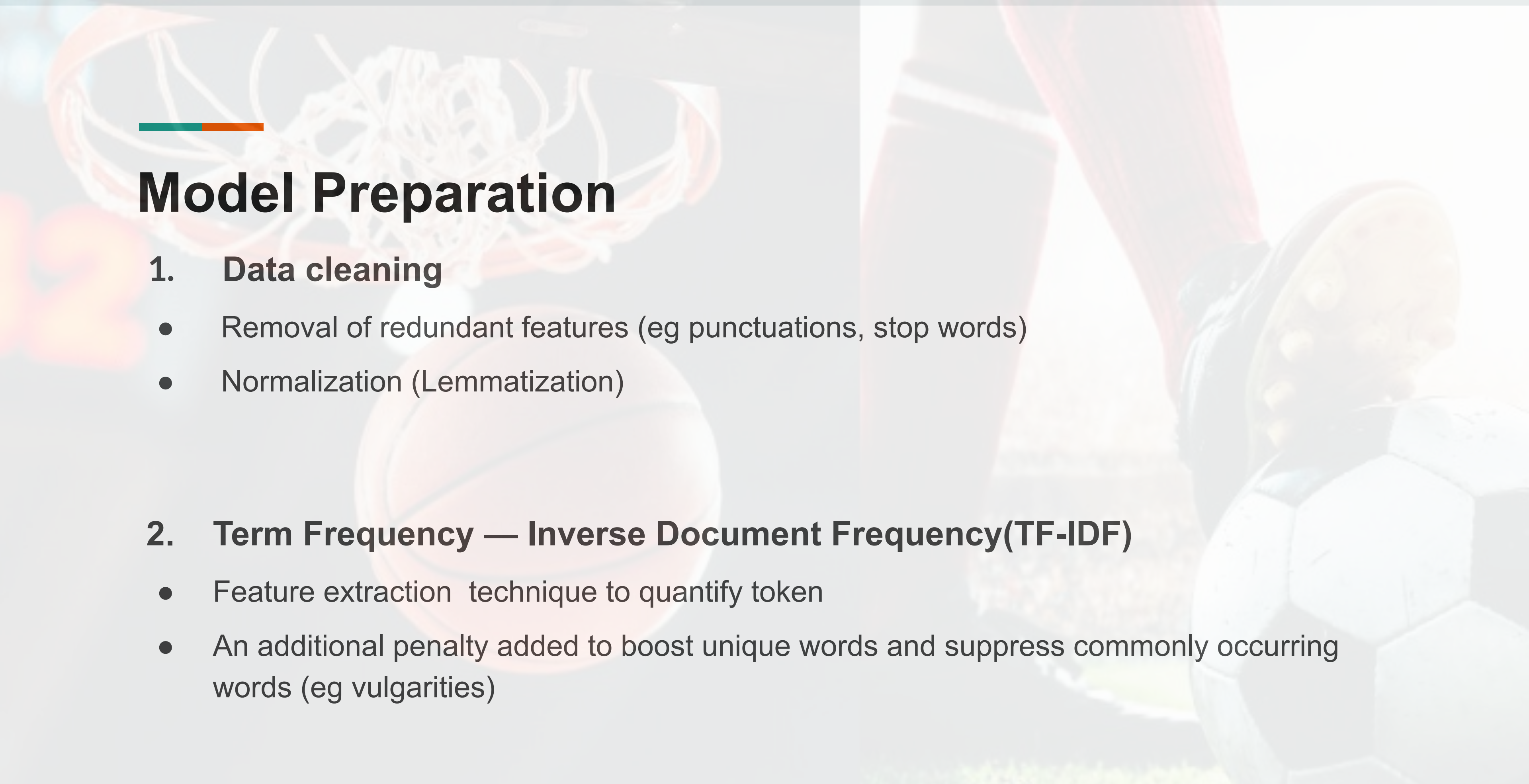
- Data was splitted into training set and testing set
- Training set was further splitted into
- Training and validation set

## Model Building

- CountVectorizer
- TfidVectorizer
- Logistic Regression
- Naive Bayes
- KNN
- SVM

## Model Selection

- Model was selected based on accuracy
- Logistic Regression
- Naive Bayes



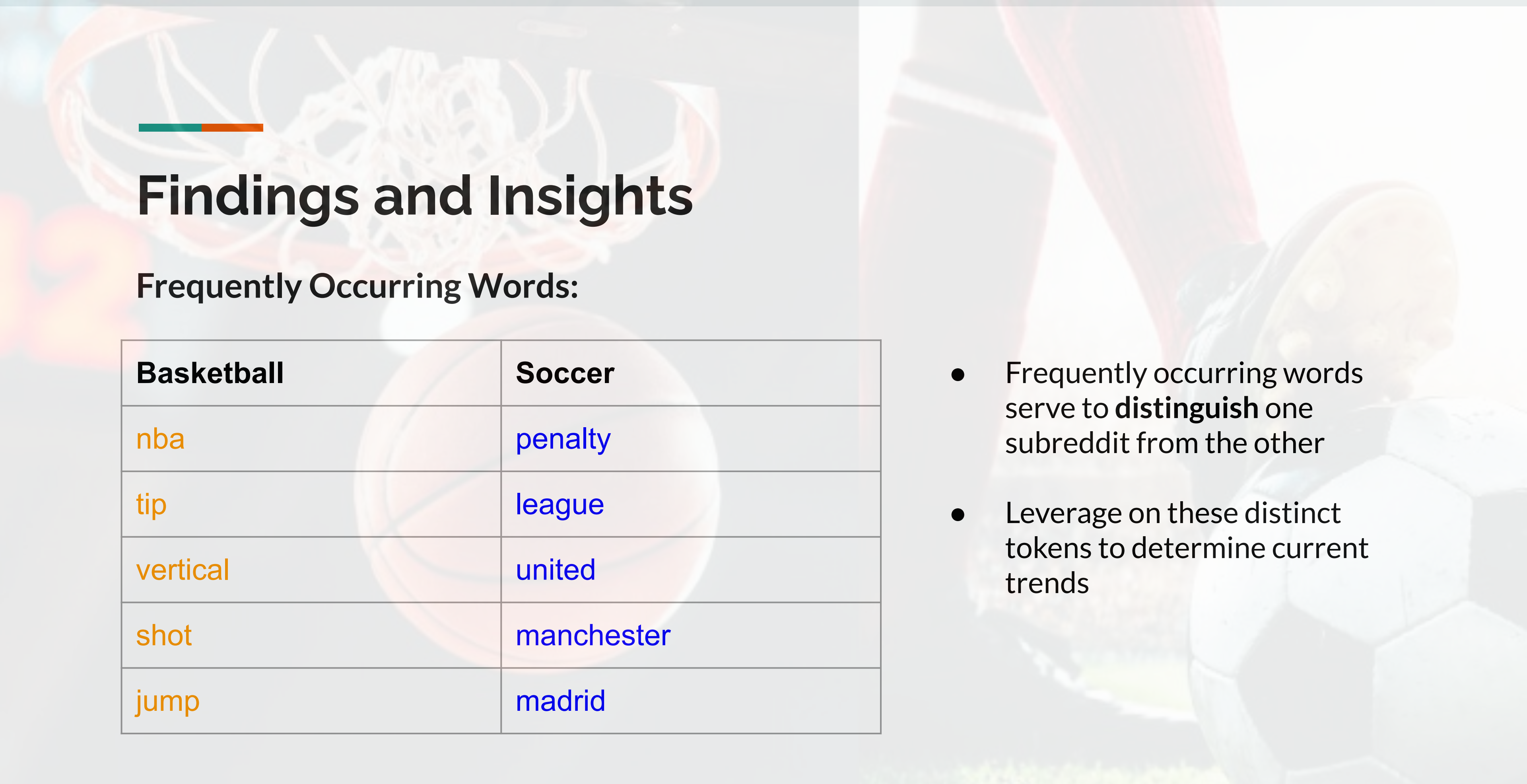
# Model Preparation

1. **Data cleaning**

- Removal of redundant features (eg punctuations, stop words)
- Normalization (Lemmatization)

2. **Term Frequency — Inverse Document Frequency(TF-IDF)**

- Feature extraction  technique to quantify token
- An additional penalty added to boost unique words and suppress commonly occurring words (eg vulgarities)

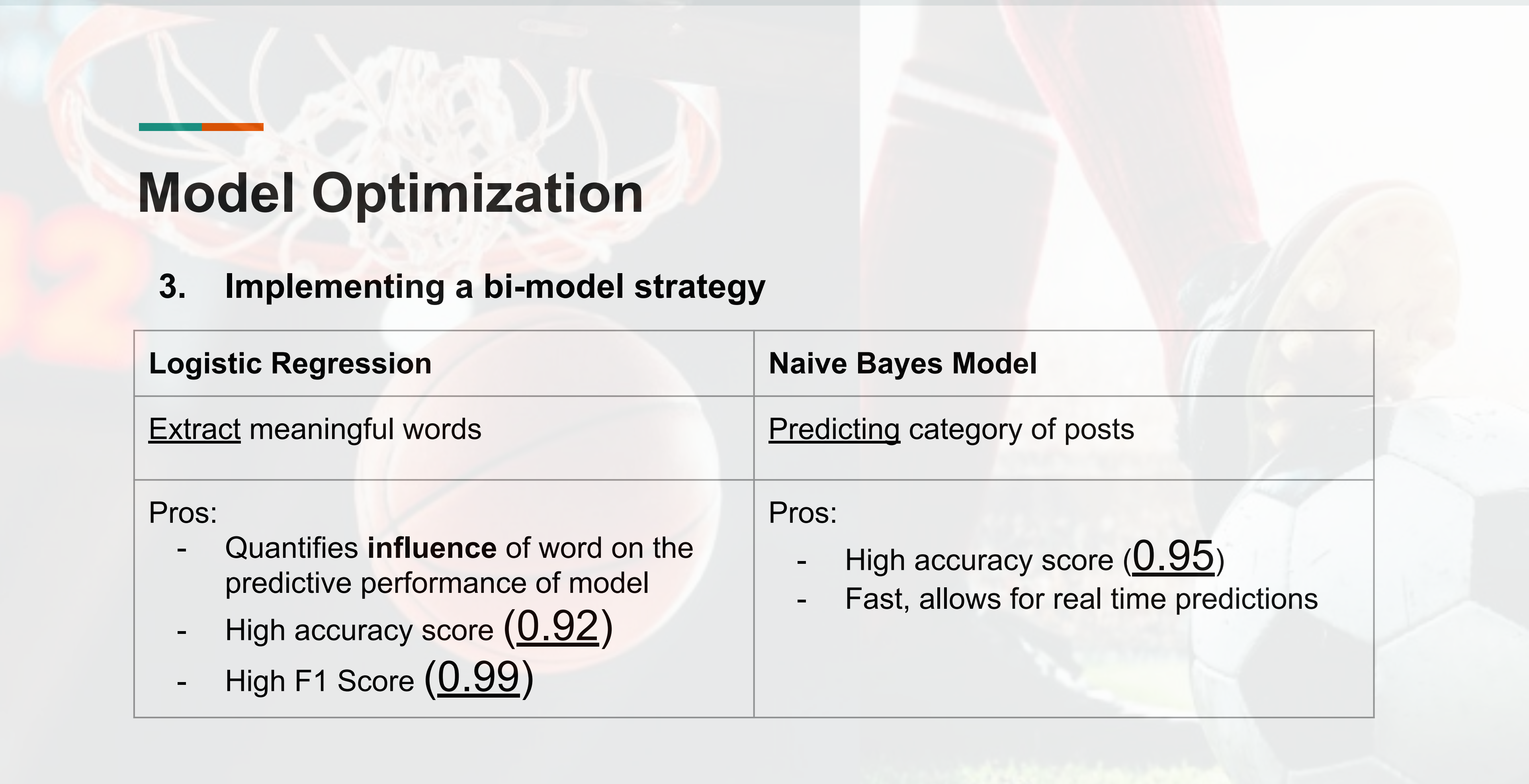


# Findings and Insights

## Frequently Occurring Words:

| Basketball | Soccer |
|---|---|
| nba | penalty |
| tip | league |
| vertical | united |
| shot | manchester |
| jump | madrid |

- Frequently occurring words serve to **distinguish** one subreddit from the other
- Leverage on these distinct tokens to determine current trends



# Model Optimization

## 3. Implementing a bi-model strategy

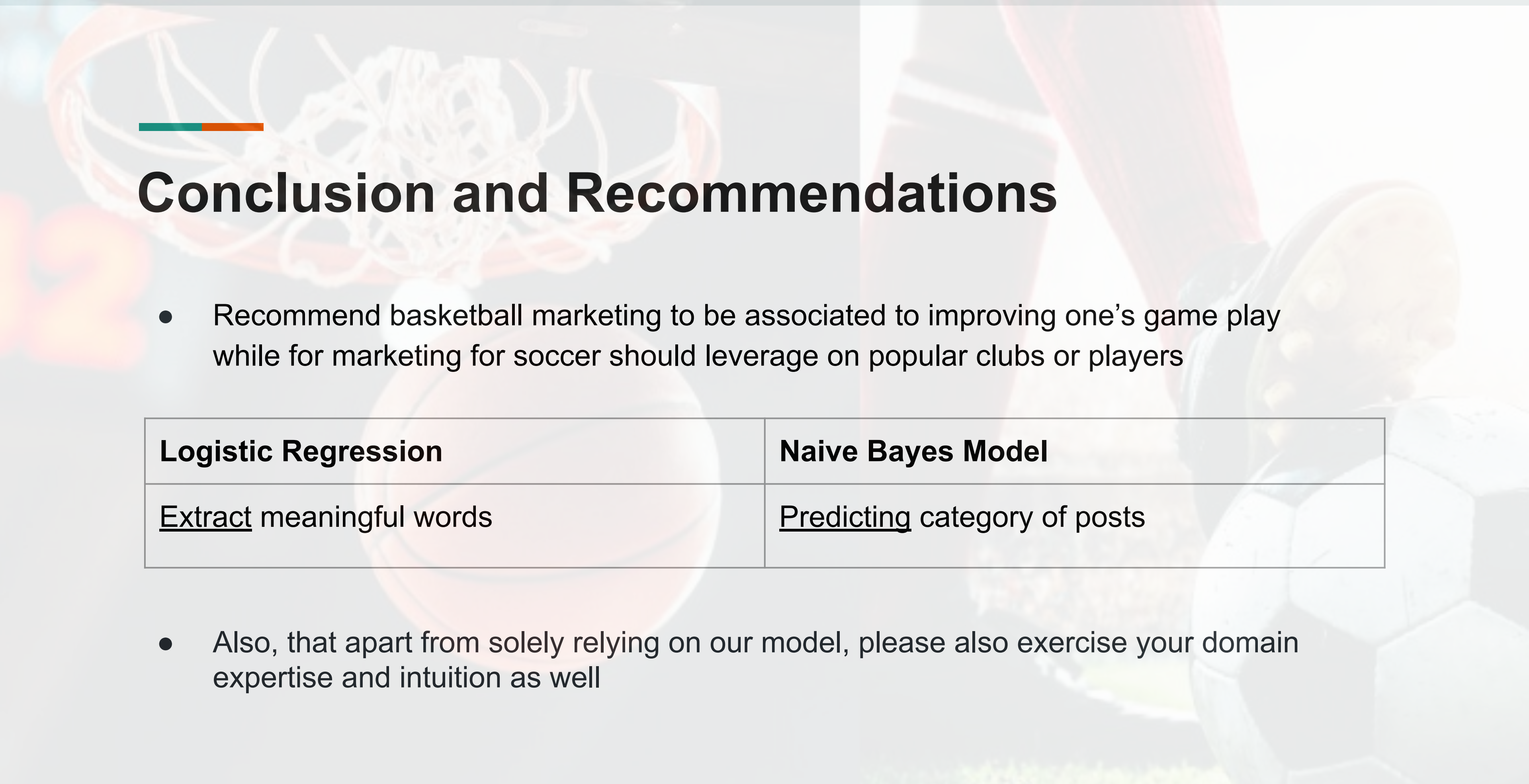
| **Logistic Regression** | **Naive Bayes Model** |
|---|---|
| Extract meaningful words | Predicting category of posts |
| Pros:<br>- Quantifies **influence** of word on the predictive performance of model<br>- High accuracy score (0.92)<br>- High F1 Score (0.99) | Pros:<br>- High accuracy score (0.95)<br>- Fast, allows for real time predictions |

# Limitations

- Our model is limited to the corpus of texts obtained from scrapping soccer and basketball reddit APIs

| Logistic Regression | Naive Bayes Model |
|---|---|
| Assumptions :<br>- Logistic model assumes linear separability between different classes | Assumptions :<br>- Naive Bayes model assumes independence between features |

- Model is only applicable to analyse basketball and soccer texts. For texts related other sports, eg american football, our model will predict the text being associated to either basketball or soccer



# Conclusion and Recommendations

- Recommend basketball marketing to be associated to improving one’s game play while for marketing for soccer should leverage on popular clubs or players

| Logistic Regression | Naive Bayes Model |
|---|---|
| Extract meaningful words | Predicting category of posts |

- Also, that apart from solely relying on our model, please also exercise your domain expertise and intuition as well

# Questions?